Paper #2

Observational Study Proposal

Weijia Li

RESEARCH QUESTION

My research interest is to find correlation between trend of Tweets referring to one company or its products and the trend of its stock price. An upward trend is defined as three consecutive rises in stock price and a downward trend is defined as three consecutive falls. I choose S&P 100 index components as samples to gather Tweets from year 2006.

The purpose of this project is trying to understand how social discussion influences market trend and to make a quantitative prediction on stock price level. If there is a strong positive or negative correlation, we can predict the level of the stock price in the next time period. If there is no correlation, I may raise another hypothesis that people do not depend on others' speculations to make their own investment and thus are always rational.

RESEARCH DESIGN

Twitter tweets will be collected as observational digital data. To reduce the noises in the sample, I will define a sub-group of users as concerned users who constantly post finance-related tweets, as they are more likely to participate in investing or other types of financial activities. Then I will analyze Twitter tweets from these concerned users to generate tweets sentiment as a proxy for investor sentiment to study the correlation between investor behaviour and movement of stock markets in two dimensions: first, how investor behavior is interacted with market macro movements and secondly, how investor behavior influences the price of a single stock. In this problem, the investors' interest to the market or a stock is approximated by the volume of tweets that are related to stock market. The outlook of the market or a stock by investors can be quantified by the sentiments of these relevant tweets.

DATASETS

Observational data will be the universe of Twitter tweets. Each data sample will be a single tweet in this universe. This data can be accessed with Full-Archive Search API for Twitter, which has instant access to the entire archive of Twitter data starting from 2006. Historical price data of stocks and indexes can be downloaded from Yahoo or Google finance within Python.

METHODS

Automated Sentiments Analysis from Twitter Tweets

Each tweet will be automatedly analyzed with natural language processing toolkits (NLKT) for its topic and sentiment and those features will be extracted to be the new data samples.

Counting Method

Number of tweets and their (positive, neutral or negative) sentiments related to the stock markets or companies over time will be counted and used as indicators of interest to invest. These counted data will later be used to forecast the recent price movement of S&P 100 index and its components.

Natural Experiment

We can easily construct a natural experiment to better study the interaction of volatile market movements with investors behaviors. The 2008 financial crisis can be seen as an “as if” random event and the tweets are always-on data stream. We can investigate the aggregated volume of tweets of different sentiments before, during and after the crisis.

Matching Method

We need to match the tweets to users to generate profiles of each Twitter user in order to determine who are more likely to participate in the financial markets than others. We can set a threshold for the percentage or absolute number of tweets a user post that are related to finance. Only users whose percentages are higher than the threshold are considered concerned users.

THE RELEVANT PROS AND CONS OF TWITTER DATA

Big

The Twitter dataset is big in ways that involves many users, each user provides enough information, and contains information in a long enough time period. Such characteristics allow us to detect small changes in a small scope and catch the big picture in the long run. 10-year tweets on the sample companies satisfies all these characteristics and will enable us to make a more precise and correct estimation.

Always-on

Twitter feeds data is always on. With this feature, we are able to study longitudinal data to investigate an unforeseen event, such as the 08 financial crisis. Since available Twitter tweets dates back to early 2006, it provides us a unique opportunity to study the concerned twitter users (needs to define this above) behaviour before, during and after

the 2008 financial crisis. Assigning the time windows, 2006 to 2007/12/1, 2007/12/1 to 2009/03/31 and 2009/04/01 to 2011/12/31, as the pre-crisis, during- crisis and post-crisis short-term investigation periods, we would be able to study the correlation of behaviors of concerned users with the situations of financial markets in the short term. We would also be able to investigate the correlation between the behaviours of concerned users with respect to a single publicly-traded company and the price movement of this company in the long term from 2012 till now.

Non-reactive

Twitter feeds data is mostly non-reactive. People can change their behaviours or even give untruthful responses when they are observed by other people including researchers. This behavior change in response to researcher measurement is called *reactivity* (Webb et al. 1966). Subjects usually are not aware of their Twitter tweets being captured and the collection of such data does not require any interaction between the subjects and researchers.

Incomplete and Non-representative

However, this dataset will be incomplete. Not all investors uses Twitter or post tweets, in fact, there may be a significant gap between actual investors and tweet posters. Investors varies by age, occupation, income level et cetera yet Twitter users have less variety. Also, not all posters are investors. Twitter is a platform that allows all people to share their information and opinion publicly and thus everyone can engage in financial discussion, but not everyone is a serious investor. One limitation of this dataset is that, we cannot verify who are the actual investors to the sample stock. However, we can define a sub-group of users as concerned users who constantly post finance-related tweets, as they are more likely to participate in investing or other types of financial activities. Searching the Twitter tweets of such users would significantly reduce the signal-to-noise ratio.

Feasibility Assessment

1. Access to Twitter Historical Tweets
    a. The Full-Archive Search API for Twitter gives you complete and instant access to the entire archive of Twitter data.
2. Method to automatically evaluate the sentiments of short text passage (NLP)
    a. Natural Language Toolkit (NLKT) is implemented in python; tutorial of how to use NLTK to collect and process Twitter data is written in IPython notebook and freely available online